IJIRR

International Journal of Information Research and Review
Vol. 05, Issue, 02, pp.5221-5225, February, 2018

RESEARCH ARTICLE

A GREEN HYBRID RECONFIGURABLE LAST-LEVEL CACHE ARCHITECTURE

[*1,2]Magdy Abdelaal

[1]Communication and Networks Engineering Department, Prince Sultan University, Riyadh 11586, Saudi Arabia
[2]Egyptian Academy of Aviation Science, Cairo airport 11776, Cairo, Egypt

ARTICLE INFO

Article History:

Received 07th November, 2017
Received in revised form
11th December, 2017
Accepted 18th January, 2018
Published online 28th February, 2018

Keywords:

Reconfigurable Hybrid Cache, Non-Volatile Memory (NVM), Chip-Multi Processor (CMP), Convex Optimization, Last-Level Cache (LLC), Spin-Torque Transfer RAM (STT-RAM), Leakage Power.

ABSTRACT

Chip-Multi Processor (CMP) designs process has been confronted by a number of serious insurmountable technological challenges such as memory wall, memory static power and the limited memory bandwidth as potential bottlenecks of the system performance. To overcome the system performance bottlenecks in future multi-core architectures, emerging memory technologies such as STT-RAM, PCRAM, eDRAM and resistive-RAM due to the many attractive features such: high density, low leakage, and non-volatility are being examined as potential replacements to existent main memories and on-chip caches. In this paper, we propose a novel reconfigurable hybrid cache configuration with hybrid memory technologies, in which STT-RAM, a type of Non- Volatile Memories, is incorporated in the last-level cache with SRAM. This reconfigurable last-level cache consists of a hybrid cache configuration and a run-time reconfiguration mechanism. Based on the read intensity or write intensity of different applications, the reconfiguration mechanism dynamically adapts the last-level cache space during run-time. We accomplish experiments on an 8-core CMP which show that the prop osed architecture achieves an average 77% and 52% energy saving over non-reconfigurable SRAM-only cache and non-reconfigurable hybrid cache architectures.

INTRODUCTION

Study has shown that over 42% of overall power dissipation in the90nm generation is consumed by leakage power (Kao *et al.*, 2002) and this value can exceed above half in 65nm technology (Kim *et al.*, 2013) (Niknam *et al.*, 2015). Therefore, in current generation processes, leakage power has become comparable to dynamic power. Also leakage power will soon exceed dynamic power in magnitude if voltage and technology are scaled down any further (Guo *et al.*, 2010). According to the significant amount of static and dynamic power consumption in recent 45nm technology the performance of microprocessors cannot continue to scale. Also due to this significant amount of power consumption in recent Nano-scale technologies, multi-core processors will not be able to afford keeping more than a small fraction of all cores active at any given moment and will soon hit a power wall by scaling (Guo *et al.*, 2010). Memory modules, on-chip storage components and specially the cache subsystem are the main components which make comparable contributions to the overall power consumption of on-chip systems. Leakage power constitutes a major fraction of power consumption of memory modules. Study has shown that “static energy is projected to account for near 70% of the cache subsystem’s energy-budget in 70nm technology” (Kim *et al.*, 2013; Muralimanohar *et al.*, 2007).

Due to the increasing trend of leakage power in future Nano-scale technologies (22nm and smaller), and also due to the substantial contribution of constituent memory-based architectures in overall power consumption of chips, architecting of new classes of memory modules and cache subsystem with the lowest leakage power is essential in these days. Emerging memory technologies, Non-Volatile Memories phenomena (NVMs), (e.g.: Spin-Torque Transfer RAM (STT-RAM), Phase-Change RAM (PCRAM), and Resistive RAM (RRAM)) with benefits such as higher density and lower leakage (near to zero) compared to the traditional SRAM- based on-chip storage architectures are potentially attractive to architect the new classes of memory modules and cache subsystem. Indeed, STT-RAM by combining of the speed of SRAM, the density of DRAM and the non- volatility of Flash memory has created a new methodology for architecting new classes of memory modules. In addition, excellent scalability and very high integration with conventional CMOS logic are the other superior characteristics of STT-RAM (Mishra *et al.*, 2011). Although the STT-RAM and other NVM memory technologies have many advantages, they suffer from a longer write latency and higher write energy consumption when compared to traditional SRAM-based architectures. In these emerging memory technologies, the write energy and write duration are higher than traditional memory technologies. To write a ‘0’ or ‘1’ , for example into an STT-RAM cell, a strong current is needed to force the storage node (Magnetic Tunnel Junction (MTJ)) reverses the magnetic direction (Hosomi *et al.*, 2005;

***Corresponding author:* Magdy Abdelaal,**
Communication and Networks Engineering Department, Prince Sultan University, Riyadh 11586, Saudi Arabia.

Zhao *et al*., 2006). In these emerging memories, the latency and energy overhead of the write operations are major obstacles in their widespread adoption (Mishra *et al*., 2011; Asad *et al*., 2017; Asad *et al*., 2015; Asad *et al*., 2017; Asad *et al*., 2016; Onsori *et al*., 2016; Safayenikoo *et al*., 2016; Niknam *et al*., 2015; Safayenikoo *et al*., 2017; Dorostkar *et al*., 2017; Sadeghi *et al*., 2017). In this paper, we propose a new reconfigurable cache architecture with hybrid memory technologies. In the proposed reconfigurable hybrid last-level cache organization, STT-RAM is incorporated with SRAM.

METHODOLOGY

Recently, in design process of Chip-Multi Processor (multi-core processor) a number of serious insurmountable technological challenges such as: memory wall, memory static power and the limited memory bandwidth as potential bottlenecks of the system performance has drawn a great deal of attention. Indeed because of these described insurmountable challenges, at 11nm over 80% of all cores may have to be dormant at all times (Karpuzcu *et al*., 2009). Also in near future, the transition from multi-core (few cores) to many-core (hundreds of cores) architectures will highlight these challenges more and more. In this section to deal with the described challenges in Chip-Multi Processors, we investigate the proposed reconfigurable last-level cache consisting of a hybrid cache configuration and a run-time reconfiguration mechanism clearly.

Hybrid cache configuration

In our proposed multicore architecture, each core has two performance counters (a read_access counter and a write_access counter). As illustrated in Figure 1, two types of memory technologies (SRAM and STT-RAM) are used by the cores in two different schemes, private and shared.

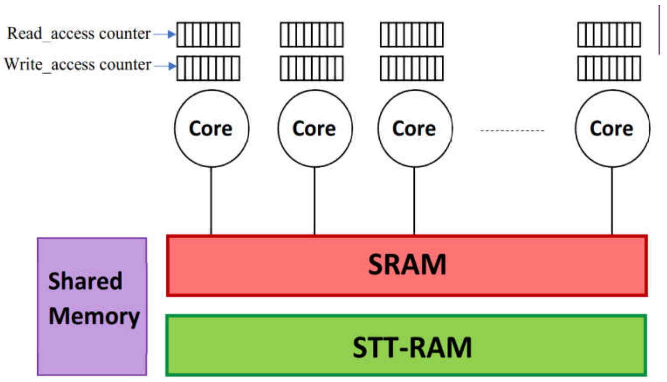

Figure 1. The proposed architecture of the reconfigurable multicore in this work

When a core accesses to its dedicated private memory segment and this access is a write access, the write_access counter counts up, otherwise, if this access is a read access, the read_access counter counts up. Since the proposed methodology is a runtime technique, it is applied periodically for the proposed multicore architecture. At the end of each time interval, based on the value of read_access and write_access counters shown in Equations (1) and (2), a suitable portion in the private part of SRAM memory or STT-RAM memory dedicated to each core is turned on.

$$\text{Equation (1)}: \begin{cases} A1:\ read_{access} \le thr_1 \\ B1:\ thr_1 < read_{access} < thr_2 \\ C1:\ thr_2 \le read_{access} < thr_3 \\ D1:\ thr_3 \le read_{access} \end{cases} \quad \text{Equation (2)}: \begin{cases} A2:\ write_{access} \le thw_1 \\ B2:\ thw_1 < write_{access} < thw_2 \\ C2:\ thw_2 \le write_{access} < thw_3 \\ D2:\ thw_3 \le write_{access} \end{cases}$$

In this context at the end of each time interval, if the write_access counter is more than a threshold shown in Equation (1), the mapped application on the core is write intensive and the number of the segments in the dedicated SRAM memory to it is increased. In this trend, if the read_access counter is more than a threshold shown in Equation (2), the mapped application on the core is read intensive and the number of the segments in the dedicated STT-RAM memory to it is increased. About the shared part of memory, we dedicate a default shared portion to all of the cores at first. Based on Equation (3), according to the number of memory intensive cores, we will increase the number of segments in the shared memory part.

This procedure is a reconfigurable approach that configures the amount of private and shared memory banks for each core according to their memory intensive or computation intensive behaviours. Based on this reconfigurable methodology with turning off the extra memory banks, we can reduce the amount of power consumption and improve performance. Since with turning off the extra memory banks, the time for searching between unwanted memory banks reduced, the performance improves significantly.

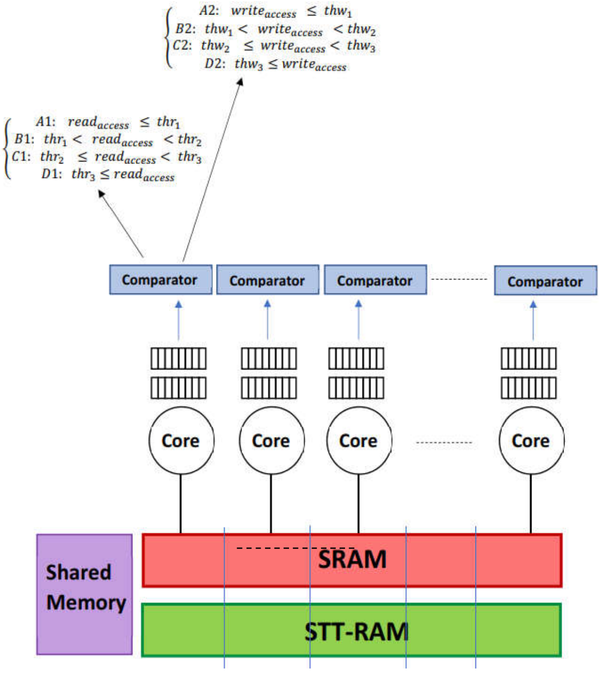
A2: $write_{access} \le thw_1$
B2: $thw_1 < write_{access} < thw_2$
C2: $thw_2 \le write_{access} < thw_3$
D2: $thw_3 \le write_{access}$
A1: $read_{access} \le thr_1$
B1: $thr_1 < read_{access} < thr_2$
C1: $thr_2 \le read_{access} < thr_3$
D1: $thr_3 \le read_{access}$
Comparator
Comparator
Comparator
Comparator
Core
Core
Core
Core
Shared Memory
SRAM
STT-RAM

Figure 2. The overview of the proposed reconfigurable approach

Figure 2 shows the overall view of this reconfigurable approach. The architecture of the proposed hybrid cache configuration is similar to hybrid cache architecture and its peripheral circuits in (Chen *et al*., 2012) as shown in Figure 3. The block diagram of the CMP system which consists of the proposed green hybrid reconfigurable last-level cache made from two memory technology regions is shown in Figure 4. In this figure, CMP system consists of multiple cores. The L1 caches are private to each core. Also in this baseline CMP system the last-level cache (L2) is shared by all the cores.

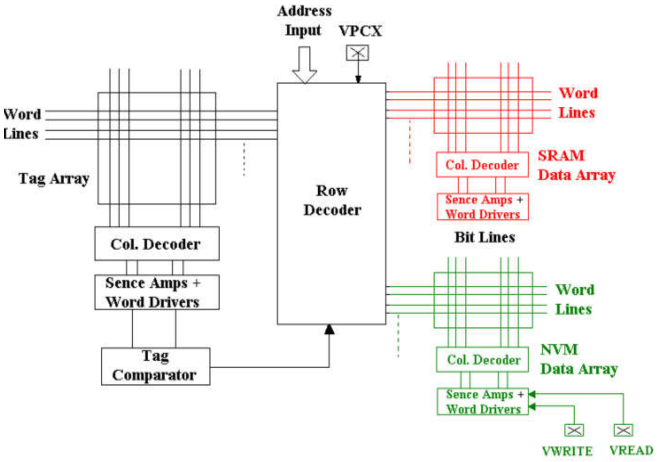

Fig. 3. The architecture details of the hybrid cache configuration (Chen *et al*., 2012)

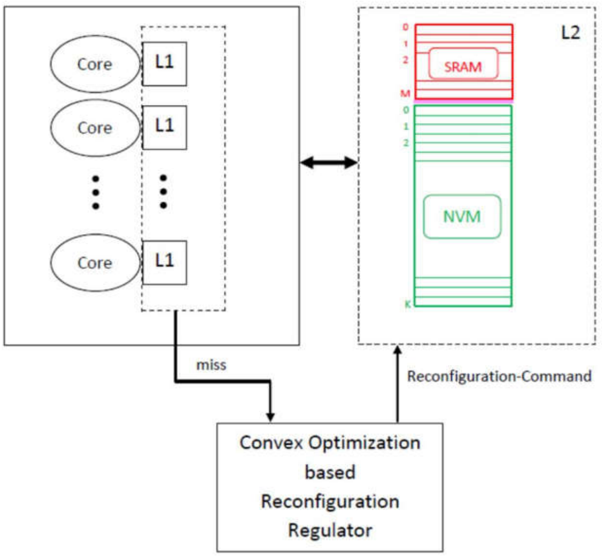

Fig. 4. Overview of the dynamic hybrid reconfigurable last-level cache

Table 1. Base line cmp configuration

No. of Cores	8
Configuration	1GHz, in-order, 14-stage pipeline
Private L1	SRAM, 64B line, size 64KB
Shared Cache	Hybrid reconfigurable cache/ SRAM+STT-RAM
Main Memory	4GB

Table 2. Energy parameters comparison at 32nm

L2 Cache Design	Read Energy	Write Energy	Leakage Power at 80^{0}C	Read at 3GHz	Write at 3GHz
1MB SRAM	0.168 nJ	0.168 nJ	444.6 mW	3 cycles	3 cycles
2MB SRAM	0.321 nJ	0.321 nJ	735.5 mW	3 cycles	3 cycles
4MB SRAM	0.622 nJ	0.622 nJ	1153.8 mW	3 cycles	3 cycles
6MB STT-RAM	0.278 nJ	0.765 nJ	319.6 mW	3 cycles	33 cycles

Run-time reconfiguration mechanism

Howbeit the proposed hybrid cache configuration provides the optimal bandwidth over the whole range of capacities, but it does not guarantee the best performance for various applications with varied bandwidth demands. Also it does not guarantee the minimum energy consumption and reliable operations for various applications. In this work a run-time reconfiguration mechanism which uses an online convex optimization- based reconfiguration regulator is used to achieve the best performance, minimize energy consumption and maximize reliability during run-time for various applications with varied characteristics.

Experiments

In this section first to evaluate the proposed hybrid reconfigurable cache architecture, the experimental setup is described. Second, different experiments are performed to quantify the benefits of the proposed architecture compared to the traditional architectures.

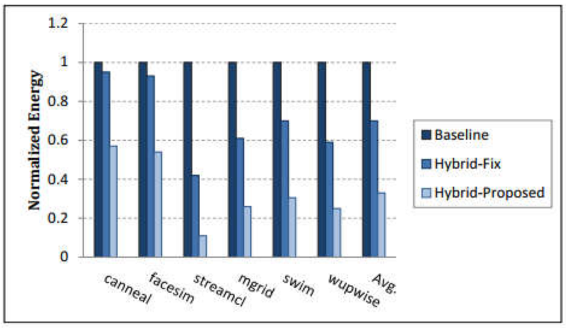

Fig. 5. Total energy comparison normalized to the Baseline

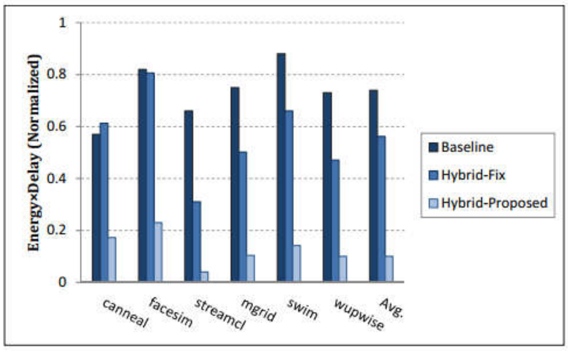

Fig. 6. Energy×Delay comparison normalized to the Baseline

Experimental setup

We evaluate our novel hybrid reconfigurable last-level cache on a simulation platform built upon Simics (Guo *et al*., 2010) to run our experiments and measure system performance improvements. The simulation platform is configured to model an 8-core CMP. Each core is in-order. We fix the SRAM-based private L1 caches to be 16K in the CMP. Table 1 shows the detailed parameters used in the CMP-architecture. We perform our evaluations based on 32nm process technology for CMOS, and STT-RAM. We simulate our proposed architecture and other design schemes on multithreaded workloads. The multithreaded applications with large working sets are selected from PARSEC benchmark suite (Bienia *et al*., 2008) and SPEC OMP2001 (Available at http://www.spec.org/omp/.). These selected benchmark suits consist of emerging workloads suitable for next generation shared-memory programs for CMPs. To model the energy of SRAM and STT-RAM memory technologies, we use the ITRS 32nm process model. Based on estimations from CACTI 6.0 (Muralimanohar *et al*., 2007), energy parameters for SRAM and STT-RAM memories used in our simulations are obtained. The energy parameters which we used in our experiments are shown in Table 2. The comparison of total energy costs is shown in Figure 5. Figure 6 shows the results of energy- product-delay improvement by our proposed technique over other two schemes. As can be seen in these figures energy consumption and energy-delay product have been improved in compared to the other baselines.

Conclusion

In this work, we propose a novel hybrid reconfigurable last-level cache architecture. This reconfigurable last-level cache consists of a hybrid cache reconfiguration and a run-time reconfiguration mechanism. Based on the predicted bandwidth demands of different applications, the reconfiguration mechanism dynamically adapts the last-level cache space during run-time. We formulate this problem as a novel online convex optimization based method that solves the reconfiguration problem to choose the high performance last-level cache configuration with minimum energy consumption during run-time. This online convex optimization method solves the reconfiguration problem based on the predicting of bandwidth demands of different applications and considering temperature distribution of cache ways to minimize the energy consumption. We evaluate the proposed design method using Simics as the simulator to run our experiments. Results show that the proposed architecture improves performance and achieves an average 77% and 52% energy saving over non-reconfigurable SRAM-only cache and non-reconfigurable hybrid cache architectures.

REFERENCES

Asad, A., Fathy, M., Jahed-Motlagh, M. R. and Raahemifar, K. 2017. Power Modeling and Runtime Performance Optimization of Power Limited Many-Core Systems Based on a Dynamic Adaptive Approach. Journal of Low Power Electronics, 13(2), 166-195.

Asad, A., Onsori, S., Fathy, M., Jahed-Motlagh, M. R. and Raahemifar, K. 2016. A heterogeneous memory organization with minimum energy consumption in 3D chip- multiprocessors. In Electrical and Computer Engineering (CCECE), 2016 IEEE Canadian Conference on (pp. 1-6). IEEE.

Asad, A., Ozturk, O., Fathy, M. and Jahed-Motlagh, M. R. 2015. Exploiting Heterogeneity in Cache Hierarchy in Dark-Silicon 3D Chip Multi-processors. In Digital System Design (DSD), 2015 Euromicro Conference on (pp. 314-321). IEEE.

Asad, A., Ozturk, O., Fathy, M. and Jahed-Motlagh, M. R. 2017. Optimization-based power and thermal management for dark silicon aware 3D chip multiprocessors using heterogeneous cache hierarchy. Microprocessors and Microsystems, 51, 76-98.

Bienia, C., Kumar, S., Singh, J. P. and Li, K. 2008. The PARSEC benchmark suite: Characterization and architectural implications. In Proceedings of the 17th international conference on Parallel architectures and compilation techniques (pp. 72-81). ACM.

Chen, Y. T., Cong, J., Huang, H., Liu, B., Liu, C., Potkonjak, M. and Reinman, G. 2012. Dynamically reconfigurable hybrid cache: An energy-efficient last-level cache design. In Proceedings of the Conference on Design, Automation and Test in Europe (pp. 45-50). EDA Consortium.

Dorostkar, A., Asad, A., Fathy, M. and Mohammadi, F. 2017. Optimal Placement of Heterogeneous Uncore Component in 3D Chip-Multiprocessors. In Digital System Design (DSD), 2017 Euromicro Conference on (pp. 547-551). IEEE.

Guo, X., Ipek, E. and Soyata, T. 2010. Resistive computation: avoiding the power wall with low-leakage, STT-MRAM based computing. In ACM SIGARCH Computer Architecture News (Vol. 38, No. 3, pp. 371-382). ACM.

Hosomi, M., Yamagishi, H., Yamamoto, T., Bessho, K., Higo, Y., Yamane, K., & Nagao, H. 2005. A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM. In Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International (pp. 459-462). IEEE.

Kao, J., Narendra, S. and Chandrakasan, A. 2002. Subthreshold leakage modeling and reduction techniques. In Proceedings of the 2002 IEEE/ACM international conference on Computer-aided design (pp. 141-148). ACM.

Karpuzcu, U. R., Greskamp, B. and Torrellas, J. 2009. The BubbleWrap many- core: popping cores for sequential acceleration. In Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on (pp. 447-458). IEEE.

Kim, N. S., Austin, T., Baauw, D., Mudge, T., Flautner, K., Hu, J. S. and Narayanan, V. 2003. Leakage current: Moore's law meets static power. computer, 36(12), 68-75.

Mishra, A. K., Dong, X., Sun, G., Xie, Y., Vijaykrishnan, N. and Das, C. R. 2011. Architecting on-chip interconnects for stacked 3D STT-RAM caches in CMPs. In ACM SIGARCH Computer Architecture News (Vol. 39, No. 3, pp. 69-80). ACM.

Muralimanohar, N., Balasubramonian, R. and Jouppi, N. 2007. Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0. In Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture (pp. 3-14). IEEE Computer Society.

Niknam, S., Asad, A., Fathy, M. and Rahmani, A. M. 2015. Energy efficient 3D Hybrid processor-memory architecture for the dark silicon age. In Reconfigurable Communication-centric Systems-on-Chip (ReCoSoC), 2015 10th International Symposium on (pp. 1-8). IEEE.

Onsori, S., Asad, A., Raahemifar, K. and Fathy, M. 2016. An energy-efficient heterogeneous memory architecture for future dark silicon embedded chip-multiprocessors. IEEE Transactions on Emerging Topics in Computing.

Sadeghi, A., Raahemifar, K., Fathy, M. and Asad, A. 2015. Lighting the Dark- Silicon 3D Chip Multi-processors by Exploiting Heterogeneity in Cache Hierarchy. In Embedded

Multicore/Many-core Systems-on-Chip (MCSoC), 2015 IEEE 9th International Symposium on (pp. 182-186). IEEE.

Safayenikoo, P., Asad, A., Fathy, M. and Mohammadi, F. 2017. Exploiting non- uniformity of write accesses for designing a high-endurance hybrid Last Level Cache in 3D CMPs. In Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on (pp. 1-5). IEEE.

Safayenikoo, P., Asad, A., Raahemifar, K. and Fathy, M. 2016. UCA: An Energy- efficient Hybrid Uncore Architecture in 3D Chip-Multiprocessors to minimize crosstalk. In Proceedings of the 9th International Workshop on Network on Chip Architectures (pp. 39-44). ACM.

SPEC OMP, SPEC OMP2001, Available at http://www.spec.org/omp/.

Zhao, W., Belhaire, E., Mistral, Q., Chappert, C., Javerliac, V., Dieny, B. and Nicolle, E. 2006. Macro-model of spin-transfer torque based magnetic tunnel junction device for hybrid magnetic-CMOS design. In Behavioral Modeling and Simulation Workshop, Proceedings of the 2006 IEEE International (pp. 40-43). IEEE.

****